US 96705

RAPIDLY QUERYABLE DATA COMPRESSION FORMAT FOR XML FILES

RELATED ARTS.
This application is 371 of PCT/IB04/52842 filed on 12/17/2004

BACKGROUND ART

The present invention relates to a method and apparatus for data compression and decompression, and particularly, to a method and apparatus for XML (Extensible Markup Language) data compression and decompression.

XML is a text format, which is becoming more and more popular in data exchange. More and more standards, e.g. multimedia field, MPEG-7 and TV-Anytime, are using XML text format to represent data.

XML is a redundant format, i.e. the way XML represents data and structures leads to a relatively large text. Therefore, data compression needs to be carefully considered for transmission or storage. The most common compression method is Zlib, e.g. the best known zip (.zip files) and gzip (.gz files). It is based on Huffman, LZ77 or both.

In the prior art, a compression device compresses the XML data and sends the compressed XML data to a decompression device, which decompresses the compressed XML data and conducts analysis therefor.

Fig. 1 is a structural diagram of a compressor in the prior art. Compressor 100 comprises LZ77 encoder 102, Huffman encoder 104 and block packer 106. Compressor 100 compresses the XML data on the basis of Zlib format.

First, Compressor 100 receives the XML data; LZ77 encoder 102 encodes the XML data according to LZ77 algorithm, generating a bunch of codewords and literals. Said literals comprise the bytes from the XML data that cannot be compressed. One codeword could convert the data previously met in the XML data, namely the redundant data, into a sequence of bytes. A typical codeword comprises length and pitch, wherein the length